A Semantic Framework for Data Quality Assurance in Medical Research

[1]Lingkai Zhu, [3]Helen Chen[*]
[1,3]School of Public Health and Health Systems
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
{[1]l49zhu, [3]helen.chen}@uwaterloo.ca
[*] Corresponding author

[2]Kevin Quach
[2]Multi Organ Transplant Institute
University Health Network
Toronto, Ontario, M5G 2C4, Canada
Kevin.Quach@uhn.ca

***Abstract* — The large amount of patient data amassed in the Electronic Patient Record systems are of great value for medical research. Aggregating research-grade data from these systems is a laborious, often manual process. We present a semantic framework that incorporates a data semantic model and validation rules to accelerate the cleansing process for data in Electronic Patient Record systems. We demonstrate the advantages of this semantic approach in assuring data quality over traditional data analysis methods.**

***Keywords* — data quality assurance, data quality measurement, ontology modelling, semantic framework, semantic web standards**

I. Introduction

Patient care is a highly complex process that involves multiple services and care providers in the continuum of care. Patient data collected may be incorrectly recorded or missing during busy clinical encounters. Thus, it is often very difficult to use patient data aggregated from a hospital's Electronic Patient Record (EPR) directly in health research which requires high quality data. Traditionally, data quality checking is performed by manual inspection and information processing, with the assistance of pre-defined data entry forms to impose data validation rules. The "cleaned" data are then stored in a research database. However, such activities must be customized to the registry platform, such as Microsoft Excel and Access. These proprietary rules are hardly interoperable with other systems and are limited in function. We propose a semantic framework that can explicitly describe the validation rules to govern data quality. The semantic framework can also perform complex cross-reference checks; whereas traditional error checking mechanisms would have difficulty incorporating, especially when the list of conditions changes over time, or changes with different application domains. Therefore, the use of a semantic framework can help accelerate and generate high quality research data over traditional techniques.

II. Literature Review

A. Categorizing Data Quality Problems

The quality of data is measured in multiple dimensions, which means "aspects or features of quality" [1]. We refer to three notable summaries of data quality dimensions [2][3][4]. Although there is no general agreement on classifications and definitions for dimensions, we identified three dimensions that are most suitable in our context: completeness, consistency and interoperability.

B. Improving Data Quality via a Semantic Framework

Brueggemann and Gruening presented three examples that demonstrate how a domain ontology can help improve data quality management [5]. According to the authors, applying semantic techniques brings advantages like suggesting candidate consistent values, using XML namespace to keep track of data origins and flexible annotation on results. We apply their three-phase methodology (construction, annotation and appliance) and demonstrate other benefits, e.g. rules expressed in semantic restrictions are more explicit than external algorithms.

Fürber and Hepp pursued a semantic approach of handling missing value, false value, and functional dependency data quality problems [6]. They chose SPARQL queries to implement rules detecting data deficiencies and described handling missing value sections that constraints, such as cardinality, are difficult to model in RDFS or OWL. However, OWL features such as owl:allValuesFrom and owl:oneOf are sufficient to model constraints from the database schema we use. We will express our semantic framework in OWL DL and SWRL. OWL DL provides class and property restrictions we need while remains decidable. DL-Safe SWRL rules are sufficiently expressive for our data quality rules, whilst provide ease of reusing already defined OWL classes and properties. This combination receives reasoning support from the Pellet reasoner[1].

III. Methodology

A. Architecture of Data Quality Assurance Framework

The data quality assurance framework is illustrated in Fig. 1 (rectangles and circles represent data repositories/ontologies and software modules, respectively). The whole framework revolves around a transplant EPR ontology, which is built with the openEHR reference model ontology [2] as the core framework, and refers to an ICD-10 ontology [3] for proper diagnoses definitions. The construction of EPR ontology starts with a script converting the database schema of an

[1] http://clarkparsia.com/pellet/
[2] http://trajano.us.es/~isabel/EHR/
[3] https://dkm.fbk.eu/index.php/ICD-10_Ontology

anonymized test medical database into an EPR taxonomy. The attributes in the database are captured in a class hierarchy and mapped into the OpenEHR ontology, and patients with data are imported as instances. Class restrictions and data quality validation rules are written in OWL and SWRL, respectively, and the Pellet reasoner handles reasoning for both. Through reasoning, data quality issues within the patient instances are recognized and annotated, which enables the data exporter module to clean the data, and provide the cleaned data to researchers for analysis.

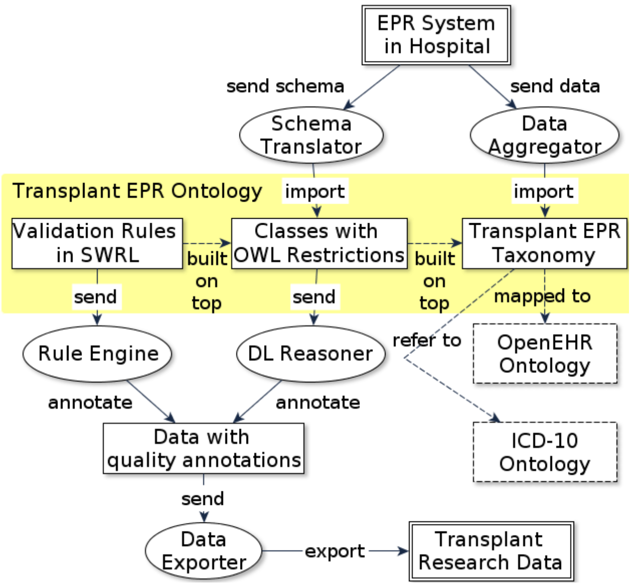

Fig. 1. Data Quality Assurance Framework Architecture

B. *Data quality assessment by dimensions*

To assess EPR data, three data quality dimensions are summarized for reference:

1. **Completeness**

Completeness refers to the proportion of data that is available in EPR relative to an expected complete dataset. This dimension can be used to examine the whole dataset as well as a single attribute.

Example: for all required attributes, instances that have at least one (by defining owl:someValuesFrom restrictions) valid value are annotated as complete.

2. Consistency

The consistency dimension refers to the logical coherence of relationships between data from different attributes, which frequently appear in an EPR domain. SWRL rules are employed to translate medical knowledge into logical connections properly.

Example: a post-transplant diagnosis cannot have a date earlier than transplant date; otherwise, it is a pre-transplant diagnosis and needs to be recorded as an error. A SWRL rule, using the date built-in, is able to identify such temporal inconsistencies and annotate them.

3. Interoperability

The interoperability dimension refers to the compatibility of a data element with other information systems. When importing diagnosis data, our data aggregator tries to seek each value in an external, standardized taxonomy, such as ICD-10. If the value is found, an owl:sameAs statement is made to map the value to the standard diagnosis definition, and the data element is marked interoperable.

IV. Preliminary Results

Restrictions and rules are implemented reflecting the identified data quality dimensions. Annotation sub-classes, such as "patient with complete demographic info", are created under the patient class. A reasoner is applied to classify all patient instances into these sub-classes. For each instance, we detect how many criteria it meets. For each sub-class, we know how many patients fall into it. Custom filters such as "patients who satisfy all rules" are also constructed. The results are manually reviewed and found correct.

V. Discussion and Future Work

Traditionally, data restrictions are enforced in an E-R database but its limited function could only ensure the completeness and the value range of data. Our semantic framework can perform the latter functions and can check for data consistency and interoperability, which brings greater benefit to medical research data quality.

The next step of our work is to repeat our methodology on a real and uncleaned EPR dataset. A research proposal has been submitted to a hospital based in Toronto with a transplant program for access to their dataset of 2000 patients. We will apply our semantic framework and identify any errors for review by researchers in the program. Once the framework's robustness and accuracy is established, EPR data in production can be checked regularly to ensure the quality of health data.

References

[1] D. McGilvray, Executing Data Quality Projects: Ten Steps to Quality Data and Trusted InformationTM. Morgan Kaufmann, 2010.

[2] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” ACM Computing Surveys (CSUR), vol. 41, no. 3, p. 16, 2009.

[3] C. Fürber and M. Hepp, “Towards a vocabulary for data quality management in semantic web architectures,” in Proceedings of the 1st International Workshop on Linked Web Data Management, 2011, pp. 1–8.

[4] P. Oliveira, F. Rodrigues, and P. Henriques, “A formal definition of data quality problems,” in International Conference on Information Quality, 2005.

[5] S. Brüggemann and F. Gruening, “Using domain knowledge provided by ontologies for improving data quality management,” in Proceedings of I-Know, pp. 251–258, 2008.

[6] C. Fürber and M. Hepp, “Using semantic web resources for data quality management,” in Knowledge Engineering and Management by the Masses, Springer, 2010, pp. 211–225.